DOCUMENT RESUME

ED 422 374 TM 028 935

AUTHOR	Ruiz-Primo, Maria Araceli; Shavelson, Richard J.; Baxter, Gail P.
TITLE	Evaluation of a Prototype Teacher Enhancement Program on Science Performance Assessment. Draft.
PUB DATE	1998-00-00
NOTE	63p.
PUB TYPE	Reports - Evaluative (142)
EDRS PRICE	MF01/PC03 Plus Postage.
DESCRIPTORS	Elementary Secondary Education; Evaluation Methods; *Faculty Development; *Inservice Teacher Education; Models; *Performance Based Assessment; Program Effectiveness; *Program Evaluation; Program Implementation; *Science Instruction

ABSTRACT

A central task in evaluating a prototype education program is to study the variability of delivery and outcomes from site to site. The evaluation should also indicate what to expect and what to do when a Teacher Enhancement Program (TEP) program becomes fully operational. An approach to evaluating TEPs was developed and applied to a concrete case, an evaluation of a prototype TEP designed to enhance teachers' knowledge and use of science performance assessments. The prototype program was implemented in two sites with different facilitators and participants. Three program components--delivery, materials, and outcomes--were evaluated successively across three iterative tryouts using multiple sources of information and multiple methods of data collection. Evaluation findings across the three tryouts showed that the program was "robust." In general, the TEPs produced similar results with different facilitators despite variations in program implementation. However, evaluation findings from direct observation revealed that during the implementation of the program, some information provided by facilitators during discussions and/or in answering participants' questions was not always accurate. Facilitators must have extensive knowledge and experience in hands-on science teaching and with performance assessments. Moreover, these facilitators must be thoroughly trained in delivering the TEP, if misconceptions are to be avoided. Appendixes list topics addressed in the programs by goal, the sequence and organization of program content, and topics and issues in the new prototype program. (Contains 6 tables, 7 figures, and 30 references.) (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 422 374

Evaluation of a Prototype Teacher Enhancement Program on Science Performance Assessment*

Maria Araceli Ruiz-Primo and Richard J. Shavelson

Stanford University

Gail. P. Baxter

University of Michigan

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Maria A. Ruiz-Primo

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

Draft

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

RUNNING HEAD: Formative Evaluation

TM028935

* The work presented in this paper was developed when the first two authors were at the University of California, Santa Barbara.

Evaluation of a Prototype Teacher Enhancement Program on

Science Performance Assessment

Abstract

A central task in evaluating a prototype education program is to study the variability of delivery and outcomes from site to site. The evaluation should also indicate what to expect and what to do when a Teacher Enhancement Program (TEP) program becomes fully operational. We presented an approach to evaluating TEPs and applied this approach to a concrete case, an evaluation of a prototype TEP designed to enhance teachers' knowledge and use of science performance assessments. The prototype program was implemented in two sites with different facilitators and participants. Three program components--delivery, materials, and outcomes--were evaluated successively across three iterative tryouts using multiple sources of information and multiple methods of data collection. Evaluation findings across the three tryouts showed that the program was "robust". In general, the TEP produced similar results with different facilitators despite variations in program implementation. However, evaluation findings from direct observation revealed that during the implementation of the program, some information provided by facilitators during discussions and/or in answering participants' questions was not always accurate. Facilitators must have extensive knowledge and experience in hands-on science teaching and with performance assessments. Moreover, these facilitators must be thoroughly trained in delivering the TEP, if misconceptions have to be avoided.

Evaluation of a Prototype Teacher Enhancement Program to Transfer Performance Assessment Technology

Current science education reform addresses fundamental questions (e.g., Hurd, 1986) such as: What teaching methods enable students to understand the nature and culture of science? How can educators foster scientific literacy in students? How can science be related to everyday decision making? How can science understanding be assessed? The reform's answers to these questions are: Science instruction should parallel the methods used by scientists to understand the natural world (e.g., Raizen, Baron, Champagne, Haertel, Mullis, & Oakes, 1989). From this perspective, students have to do science--observe, hypothesize, record data, draw inferences and make generalizations--to solve scientific problems. By "doing" science students construct meaning both individually and in groups. Finally, assessment of student learning should parallel instructional reform.

Unless current assessment practice is changed, however, assessment will not parallel instruction reform, and reform in science education will not be comprehensively implemented in the classroom (e.g., Kulm & Stuessy, 1991; Shavelson, Carey, & Webb, 1990). Consequently, many states have responded with new policies which move achievement testing away from multiple-choice tests of basic skills toward performance-based assessments of knowledge and problem solving.

Changes in the nature and purpose of science instruction and subsequent changes in the nature of assessment exert pressure on the classroom teacher to change instructional and assessment practices (e.g., Shavelson & Baxter, 1990). Teachers are expected to shift from textbook and rote memory to constructivist teaching--teaching based on students' active

construction of knowledge in problem-solving situations. To teach this way, teachers need to be well grounded in science to support an inquiry approach. They also need to change their role in the classroom from conveyors of facts and concepts to facilitators of knowledge construction. Moreover, they need skills in managing the physical and social organization of the classroom to support inquiry teaching (e.g., small groups of students working together). Finally, teachers need to have knowledge about new assessment policies and practices.

To support teachers in the transition from traditional textbook teaching to constructivist teaching, a sustained program of in-service education is needed. Such a program would give teachers an opportunity to deliberate about the new perspectives in curriculum, teaching, learning, and assessment (e.g., Hurd, 1986; Shavelson, Copeland, Baxter, Decker, & Ruiz-Primo, 1994). In response to this need, the National Science Foundation (NSF) initiated in 1984 a Teacher Enhancement Program (TEP) to provide effective in-service education and foster the development and dissemination of improved models for conducting in-service education programs for science and mathematics teachers across the country (e.g., Fitzsimmons, Carlson, Burnham, Heinig, & Stoner, 1991).

One of the main goals of agencies like NSF is to have prototype in-service programs with significant impact not only on the sites where the program was developed but also in other settings (e.g., Fitzsimmons et al., 1991). Consequently, information about how these prototype programs were developed, what their characteristics are, and how they can be transferred to other sites has been requested not only by agencies like NSF but also by Congress (e.g., Fitzsimmons et al., 1991; Knapp, Shield, St. John, Zucker, & Sterns, 1988).

The purpose of this study was to conduct a formative evaluation of a prototype TEP which aims to provide teachers with the knowledge and skills needed to understand, use, and select science performance assessments. The formative evaluation tested the prototype program at different sites with a variety of facilitators and a variety of participants. It provided information about program components that need adjustment and revision to increase the probability of the program succeeding at different sites. Here we report on the transfer of a prototype TEP from the development site to outside sites.

Formative Program Evaluation

The practice of program evaluation has, once again, caught the attention of policy makers. They recognize that there is limited knowledge about the design and development of successful TEPs necessary for the reform of science education. What is needed is an understanding of the process involved in the development and implementation of successful in-service programs. Simply put, evaluative information is vitally needed.

Formative evaluation helps program developers better understand how and why the program is a success or a failure, to specify what aspects of the program are relatively more successful than others, and among which groups of participants (e.g., Cronbach et al., 1980). The main goal of formative evaluation is to modify and improve the design of any program while the program is still under development (e.g., Scriven, 1967) and therefore capable of being revised. Formative evaluation provides data as a basis for improving in-service programs (e.g., Chinien & Hlynka, 1993), helps to make judgments about how successful the program is (e.g., Guskey & Sparks, 1991), and helps to accumulate knowledge about how effective in-service programs are developed and adapted.

In spite of a general consensus among the policy makers and practitioners regarding the importance of formative evaluation, the great majority of in-service programs are still being implemented without prior formal evaluation (e.g., Knapp et al., 1988). If any kind of evaluation is carried out, it usually focuses either on changes in tests scores, or on information that is likely to be of tangential interest and utility to knowing and learning more about effective programs (See Ellis & Goulding, 1991; Ellis & Kuerbis, 1991; Gayford, 1987 as notably exceptions). The quality of in-service programs needs to be investigated, from planning and designing of the programs to follow-up of their impact.

The information provided from systematic formative evaluation may be used by developers and administrators for program improvement. Funding agencies and policy makers (e.g., NSF, Congress) may use the information to assist program developers with ways to formatively evaluate their program. For example, they can use the information to consider the implications of the evaluative information from any one in-service program for others also being operated by the same agency; or to disseminate how successful programs were developed, which may help other principal investigators who are leading the development or implementation of in-service programs.

An Approach to Formative Evaluation of TEPs

The approach proposed by Ruiz-Primo (1994) was used to carry out the evaluation of the prototype TEP. This approach makes two important assumptions. The first assumption is that a central task of evaluation is to facilitate the transfer of knowledge from some programs or sites to other programs or sites by explaining the processes that lead to the outcomes

achieved (e.g., Cronbach, 1982.) The second assumption is that formative evaluation seeks to provide information to improve program performance by influencing immediate decisions about the program, especially about how its component parts and processes could be improved (e.g., Scriven, 1967, 1991a, b; Shadish, Cook, & Leviton, 1991).

This approach captures two types of information: (a) information related to the intrinsic value of the in-service program, and (b) information related to its potential dissemination. Both types of information help decision makers adjust and improve the program; however, each provides information on different aspects. Information about the intrinsic value refers to whether or not the program components (e.g., context, materials, and delivery conditions) are likely to meet the program goals. Information on the potential dissemination refers to how generalizable the program is to other settings (e.g., Weiss, 1972).

This approach to formative evaluation is built on three major elements: (a) the characteristics of the TEP to be evaluated; (b) the process involved in conducting the formative evaluation; and (c) the role and knowledge of the evaluator carrying out the formative evaluation (see Table 1).

Insert Table 1 Here

TEP Characteristics. A TEP can be characterized as a system of interrelated *components*--context, goals, materials, delivery, and outcomes--which develop through three *stages of maturity*: (1) the *"planned program"*--the turn of an idea into a program for action; (2) the *"experimental program"*-

-a trial program to see what the program can accomplish, and (3) the *"prototype program"*--a model program that attempts to preview what will happen when the program is fully operational.

Evaluation Process. The formative evaluation process is conceptualized as an *iterative process* in which the program's goals are realized through successive approximations. The characteristics of the iterative process vary with the TEP's stages of maturity: from program reviews and revisions at the planned-program stage to program tryouts at different sites at the prototype program stage.

Congruent with the stages of maturity and the variations across successive trials (iterative process), the approach proposes variations in the evaluation process across the three *stages of formative evaluation*: (1) *"in-house reviews"* in which the evaluation provides information on the accuracy and adequacy of the planned-program materials, and on how feasible its operation is; (2) *"in-house tryouts"* in which the evaluation provides information on how the experimental program operates with typical facilitators and participants and what factors are associated with the program's success; and (3) *"outside tryouts"* in which the evaluation tests the prototype program in circumstances and with the population that matched intended use when the program is fully operational. The approach stresses the use of different methods (quantitative and qualitative) and sources of information in the formative evaluation. These methods and sources vary as to the stage of formative evaluation, the information needed, and the audience requesting the information.

Evaluator's Role. Finally, for the formative evaluation to achieve its goal of improving a TEP, the approach assumes that the evaluator (1)

possesses an *extensive knowledge* of the content of the program evaluated, and (2) is able to *adapt his/her role* during the evaluation.

Comment. The approach recognizes that the development stages are not necessarily linear. There is always the possibility that at the experimental or prototype stage some components may have to return to a previous stage. In this way different components of the program may be at different stages of maturity at the same time (e.g., Cronbach et al., 1980). For example, during the evaluation of a prototype program, some activities may prove to be effective under all conditions, others may need minor adjustments while still others may have to be eliminated and new activities included to achieve the TEP's goals.

The Program Evaluated: A TEP to Inform about Science Performance Assessments to Teachers and Other Educators

As a part of a project funded by NSF (Shavelson & Baxter, 1990), a team at the University of California, Santa Barbara and the University of Michigan is in the process of developing two TEPs to transfer performance assessment technology to teachers and other educators. This section describes the characteristics of the TEP that were the focus of this formative evaluation.

The TEP is part of a larger project (Shavelson & Baxter, 1990) devoted to: (a) capturing the new technology involved in developing science performance assessments; (b) providing teachers and other educators with the knowledge and skills needed to understand, select, and use performance assessments embedded within the curriculum; and (c) training teachers and other educators to create and evaluate performance assessments.

The project is organized in an overlapping sequence of three Phases: Performance Assessment Technology, Training Development, and Field Test

(Figure 1). In Phase I, Performance Assessment Technology, the emerging technology of creating performance assessments, is studied. The goals of this phase are to produce and evaluate performance assessments to be used as part of teacher pre-service and in-service education, and to make explicit the new technology's concepts and procedures so they can be transferred to teachers and other educators.

Insert Figure 1 Here

In Phase II, Training Development, a two level system of teacher enhancement is being developed. Level I training provides pre- and in-service teachers with knowledge and skills to understand, select, and use performance assessments. Level II training provides district personnel, teachers, scientist and other educators, working in teams, with the knowledge and skills needed to create and psychometrically evaluate alternative assessments. The major activities in Phase II are the development and evaluation of Level I and Level II training, culminating in prototype programs for field testing.

Phase III, Field Test, involves field testing both the Level I and Level II prototype programs to evaluate how well the training can be implemented in school districts with hands-on elementary science curricula, and the degree to which training meets its goals.

The project has two Principal Investigators with extensive experience in the development and evaluation of science performance assessments. For the development of Level I training the project has two instructional developers with extensive experience in developing teacher enhancement

programs. The head of the development team is an expert in adult education. Level I training has also a Coordinator who oversees all the activities related to the development of the program.

During the development of the TEP for Level I training a formative evaluation was carried out following the approach described above. This paper focuses on the third stage of development--outside tryouts with the prototype program over a nine-month period.

A program is becomes a prototype when the best possible program has been put together to meet its goals. The program can then be tried out in different sites, with different facilitators, and different participants. This stage may reveal a range of possible problems in operating the program on a large-scale (dissemination problems) and/or the components of the program that need to be modified in their delivery to improve effectiveness.

Characteristics of the TEP. The goals of the TEP are to provide pre- and in-service teachers with the knowledge and skills to: (1) understand the nature of assessment reform, (2) use these assessments in their classrooms, and (3) select existing assessments that are appropriate for evaluating individual student achievement or for monitoring the curriculum (Shavelson & Baxter, 1990). The mechanism for realizing these goals was a *prototype program* package that could be "exported" to school districts and be used by trainers (science educators) in those districts.

The TEP can be characterized as a *training approach* to staff development (Sparks & Loucks-Horsley, 1990). First, it is a workshop-type program in which the facilitator is the expert who establishes the content and flow of activities. Second, the training sessions are conducted with a clear set of objectives for learner outcomes. Third, the facilitator's role is to set the activities that will aid teachers in achieving the desired outcomes. This

training approach is considered useful for realizing outcomes such as awareness, knowledge, and skills development, or when teachers require demonstrations of and practice on instructional techniques to be able to use the skills in their classrooms (e.g., Joyce, 1988; Joyce & Showers, 1980; Spark & Loucks-Horsley, 1990).

The Experimental Program. The experimental program stage of development was evaluated over five tryouts with project staff as facilitators and elementary science teachers, from the Science for Early Educational Development Project (SEED) in Pasadena, CA, as participants (see Ruiz-Primo, 1994; Ruiz-Primo, Baxter, & Shavelson, 1993). The evaluation collected information from a number of sources--documents, developers, participants, facilitators, evaluator--using a wide variety of methods--review of documents, direct observation, participants' products, responses to questionnaires, and interviews with facilitators.

The in-house evaluation revealed that participants acquired information they perceived as helpful to understand, use, and select performance assessments. The magnitude of the pre- and posttest program differences increased as the TEP's materials and delivery were improved based on the formative evaluation findings across the tryouts. The evaluation also revealed that facilitators' knowledge and experience in administering and scoring performance assessments were critical to the program achieving its goals (Ruiz-Primo, 1994; Ruiz-Primo, Baxter, and Shavelson, 1993).

These evaluation findings impacted the project in at least two ways. First, the original plan to "export" the prototype program as a package to other school districts was considered unrealistic. Second, it was clear that systematic training for facilitators was needed.

The Prototype TEP. After the five tryouts and many revisions of the program, the prototype TEP, the one evaluated here, had the following characteristics:

(a) The TEP's goals were three: understanding, use, and selection of performance assessments.

(b) The program reflected a hands-on instructional approach. Participants carried out hands-on elementary science performance assessments. With three of the assessments, they conducted the investigation and scored performance using procedure-based, evidence-based, and rubric (holistic) scoring systems. They conducted an exercise on interpretation of performance assessments scores, and another on selection of performance assessments. Figure 2 shows schematically the content for each goal.

Insert Figure 2 Here

(c) The TEP addressed 18 topics nested within one of the three goals: understanding (6 topics), use (5 topics), and selection (7 topics) of performance assessments. (Appendix A presents the topics by goal.)

(d) The program package included the *"Facilitators' Manual"*, the *"Participant Notebook"*, *"Nine Elementary Science Performance Assessments"*, *"Transparencies"*, and *"Videos"*. The Facilitators' Manual is a detailed written script with the content, activities, and plan for delivering the program. For delivery purposes, the sequence and organization of content and activities were divided in "segments" (i.e., units, see Appendix B). The Participants' Notebook provides reduced

copies of the transparencies used during the program and space for recording their notes and thoughts about each segment. Participants keep it for future reference.

(e) The program was designed to be delivered in 15 hours over three-days.

The Formative Evaluation Process

The stages of the formative evaluation of the Level I training program are presented in Figure 3. This study focused on the third stage: outside tryouts. The prototype TEP was tested in two sites with different facilitators and participants.

Insert Figure 3 Here

In this stage, the formative evaluation provided information on the adaptations needed to increase the probability of success when the program is fully operational. A central evaluation task, then, was to study how delivery and outcomes varied from site to site. Since the reproducibility of program results in different sites depends, in part, on how well the enactment of the TEP is described (e.g., Cronbach, 1982), evaluation findings also focused on identifying how the variations observed across sites were related to the characteristics of the program material and how these variations might be narrowed by adapting program materials.

With this perspective the evaluation of the prototype TEP focused on three components: *delivery, materials, and outcomes* (see Table 1). Program *delivery* refers to the conduct of the program with participants--how the content is conveyed to or constructed with the participants during the delivery. Program *materials* includes all documents that describe the program's content and activities, the sequence and organization of content

and activities, and the delivery plan. Program *outcomes* refers to the participants' knowledge and skills about performance assessments acquired in the program.

Formative Evaluation Questions. The evaluation asked the following questions: (1) *Delivery*--Was the program delivered as it was designed so the program's goals can be achieved? (2) *Materials*--Which aspects of the materials led to major inaccuracies or variations during the delivery? and (3) *Outcomes*--Were the program's outcomes different from those found in previous tryouts?

This evaluation, then, focused on whether the prototype program was implemented as expected in other sites with the same effects as those obtained where the program was developed.

The evaluation of program *delivery* centered on the characteristics of the "facilitators" and the enactment of the program's instructional methods. Information was collected on the: (a) facilitators' knowledge of program content--how accurately was the content delivered; and (b) implementation of the instructional plan--how adequate was the implementation.

Program *material* was evaluated as to how the characteristics of the content and the activities contributed to variations in the implementation of the program. Information was collected on the facilitators' perceptions of the program materials--content and activities and the instructional plan to deliver them.

The *outcomes* evaluation focused on program goals: to provide participants with the opportunity to "become familiar" with, not "experts" in, the nature, use and selection of performance assessments. Information was collected on the participants acquisition of knowledge about performance assessments.

Formative Evaluation Design and Instruments. The formative evaluation design followed from the Approach to the Formative Evaluation presented previously. Three outside tryouts were carried out at two sites each viewed as an iterative pilot study. This iterative process provides cumulative knowledge about the program which increases the program's robustness (e.g., Berk & Rossi, 1990).

For each tryout, the evaluation design called for collecting information before, during and after program delivery. The evaluation, then, took place before and after the delivery of the program, as well as in a pretest-posttest design during the delivery.

To provide a comprehensive view of the program as well as to cross check findings, different sources and methods of data collection were used. Evaluation data were collected from four sources: (1) documents, (2) facilitators, (3) participants, and (4) evaluator. Three data collection methods were used: (1) direct observation, (2) questionnaires, and (3) review of documents.

Table 2 presents a schematic representation of the formative evaluation design. This Table shows, for each component of the program evaluated, the sources of information, the instruments used to collect the data, and the point in time at which the instruments were administered during the tryout.

Insert Table 2 Here

The evaluation of the program *delivery* used the evaluator as the main information source, and direct observation of the delivery as the main data collection method. To examine the program delivery, direct observation data

were collected on each program topic. The delivery was videotaped and field notes were taken. The field notes included (a) time, (b) activity, (c) comments, and (d) suggestions. They were primarily descriptive, although the evaluators' reflections, interpretations, and direct suggestions made during the observation were also noted. Textual quotations were rarely included, paraphrasing was more typical. Participants served as a secondary information sources, responding to an "Opinion Questionnaire." This questionnaire used a Likert-type rating scale that elicited participants' perceptions about important topics and activities in the program. It also included open-ended questions that asked their opinion about the content and organization of the program.

Program *material* was evaluated using documents and facilitators as the main sources of data, and review of documents and questionnaires as the main data collection methods. Participants served as a secondary information sources to cross-check the findings of the other two sources. The evaluation of the material focused on the characteristics of the content and activities that allowed for variations on the implementation of the program by facilitators other than the project staff.

The Facilitators' Manual was considered the main document to be reviewed because it contained all aspects of the program: the sequence and organization of content and activities for the three days. These reviews also focused on the accuracy of the content and the adequacy of the instructional plan for delivering the content.

At the end of each tryout information on program material was collected from facilitators and participants. Facilitators responded to the "Facilitators' Critique Questionnaire" about the content and the activities of the program. The questionnaire used a Likert-type rating scale and open-

ended questions that asked their opinion about the content and organization of the program material, and recommendations for changes in delivery, and the program as a whole. Participants' answers to the "Opinion Questionnaire" were used as a secondary source of information when necessary.

Program *outcomes* were evaluated using participants as the only source of information, and questionnaires as the method of data collection. The participants' knowledge of the content was evaluated by the "Self-Report Knowledge Inventory" in a pretest-posttest design. This inventory is a self-rating questionnaire that provides information about participants' knowledge of major topics covered in the workshop. Even though this type of instrument is not an achievement test, it has been shown to correlate highly with actual achievement, takes only a short time to administer, and is not threatening to teachers (see Tamir & Amir, 1981; Young & Tamir, 1977). The "Opinion Questionnaire" completed by the participants at the end of each workshop was also used as a secondary source of information.

Instruments were revised from one tryout to the next on the basis of their psychometric properties (when possible). Face validity was the criterion used to evaluate the validity of the participants' Self-Report Knowledge Inventory, and the Facilitators Critique and Opinion Questionnaires. Face validity--"the extent to which an instrument looks as if it measures what it is intended to measure" (Nunnally, 1970, p. 149)--is considered one of the best ways to facilitate decision makers' understanding of and belief in evaluation data (e.g., Patton, 1984). The instruments were revised by the developers and the principal investigators to increase their face validity (see Scriven, 1991a). Changes and adaptations were made to the instruments on the basis of developers' and principal investigators' comments, and the characteristics of

the content. For example, new items were included in the Self-Report Knowledge Inventory when new topics were included in the content of the workshop.

Reliability was indexed by internal consistency. Reliability coefficients for the Self-Report Knowledge Inventory were obtained at both pretest and posttest on each tryout. Coefficients are presented when evaluation findings on the workshop outcomes are discussed.

Characteristics of the Outside Tryouts. Table 3 summarizes the general characteristics of the three outside tryouts. The characteristics include Facilitators' characteristics, participants, incentives for participation, duration, and data collection methods.

Insert Table 3 Here

The program was piloted at two sites: Site 1--Southern California Superintendent of Schools, and Site 2--Middle Arizona School District. At Site 1 the program was piloted on two occasions with one elementary and one high school teacher as facilitators on both occasions. Participants were in-service teachers from different school districts in the county. At Site 2, the program was piloted once with two science resource teachers and one elementary teacher as Facilitators. Participants were in-service teachers and resource persons from the District.

At Site 1, participants paid a fee to the County Office to participate and get one unit credit course. At Site 2, participants were given time off from work. The first outside tryout, Site 1 occasion 1, was carried out in a three-day session. The second tryout was carried out in two evenings and one full-day. The third tryout at Site 2, was delivered in two full days.

The only data collection method not used on the first tryout was the Facilitator Critique Questionnaire. However, information was obtained from in-depth interviews carried out with both facilitators by one of the developers of the program. For the two remaining tryouts, all sources of information listed in Table 3 were used.

Facilitators Characteristics. Table 4 presents information on the characteristics of the facilitators. Facilitators from Site 1 were selected by the County Superintendent's Office. Both were females with 8.5 years experience, on average, as teachers and 3 as facilitators on different in-service programs. They felt they had adequate experience in hands-on science teaching and performance assessment, although information from direct observation during the delivery revealed that this might not be the case. Facilitator 1's background was education (i.e., elementary education and curriculum and instruction), whereas Facilitator 2 clearly had a science background (i.e., Zoology and Biological Science).

Insert Table 4 Here

Both facilitators were informally "trained" on the two occasions by one of the program developers. One of the facilitators, F1, experienced the program as a participant in a previous implementation delivered by the project's staff. Facilitators received the Facilitators' Manual about two weeks before the tryout. Then, facilitators and the developer met to discuss the content and the logistics of the program for approximately 16 hours before the delivery. The developer walked them through the program (e.g., the sequence, the activities, implementation, use of transparencies) based on previous implementations. The developer also answered questions the

facilitators had about the content. Finally, the developer met with the facilitators at the end of each session during the implementation. At these meetings, developer and facilitators discussed the content to be delivered during the next session.

Facilitators at Site 2 were males. They worked at a School District well recognized for its science curriculum and hands-on approach to teaching. Two of the facilitators, F1 and F2, have been resource teachers in the Districts' Science Resource Center for approximately 17 years. They had extensive experience as hands-on science teachers (23 year on average) and as trainers of teachers in the same district (9 years, on average). They also had experience in administering and scoring performance assessments. Facilitator 3 was chosen by the District for two reasons: They wanted him to be involved in the development of performance assessments for the Resource Center's curriculum units, and to be a trainer of teachers in the District on the use of performance assessments.

These facilitators were also informally "trained" by the Coordinator of the Level I training project. The three facilitators were participants in two previous implementations of the program. The first was delivered by the project's staff (see above) eight months before the tryout. The second was specifically arranged to prepare them to facilitate the training at their home site. They had the Facilitators' Manual and met with the Coordinator for approximately 6 hours to discuss program content and a plan to deliver it. During the implementation of the tryout, the Coordinator also met with the facilitators at the end of the first day to discuss concerns about the content to be delivered at the next session.

Participant Characteristics. Table 5 presents participants' characteristics across tryouts. Participants differed from site to site. Whereas Site 2 has been

recognized as an exemplary hands-on school district, Site 1 is starting to move to a hands-on instructional approach. At Site 1, some teachers in some school districts were already using this approach, while others were not familiar at all with this new way to teach science.

Insert Table 5 Here

Most participants at both sites were elementary teachers. A few of them, particularly at Site 2, were junior high and high school teachers. Forty percent of the participants held a master's degree, 26 percent in education. Only one participant held a Ph.D. in education.

Evaluation Findings

This evaluation focused on three major questions about the prototype program: (1) "Is the program delivered as it was designed so the program's goals can be achieved?"; (2) "Which characteristics of the program's material lead to major variations across the sites?"; and (3) "Do the program's outcomes differ from those found in previous tryouts?"

Data were brought to bear on each program component--delivery, materials, and outcomes--for the three outside tryouts. To examine program delivery and materials, data were collected on each of the topics that constituted the program (see Appendix A). Data bearing on the outcomes component were based on pretest-posttest scores from the Self-Report Knowledge Inventory and the participants' opinions.

First, a summary of the evaluation findings across tryouts is presented along with the major decisions made about the content and delivery of the

program. Next, two examples of the evaluation findings on program delivery and program material are presented.

Summary of Evaluation Findings

From revisions made to the experimental program during the five tryouts of the in-house evaluation, the TEP program was considered to be ready for implementation at other sites--the content and the instructional plan were adequate and the program had proven effective in achieving its goals. However, the evaluation findings that emerged through the outside tryouts revealed that the program still needed some adaptations to increase the likelihood of successful transfer to other facilitators at other sites. These findings are presented for each of the components evaluated.

Program Delivery. Evaluation during the delivery of the program produced four major findings: (a) The delivery of the program was modified by facilitators on all three tryouts. Modifications were: (1) "Superficial"--facilitators added new activities not directly related to the content or the goals of the program (e.g., to give prizes to participants, play games). (2) "Process"--facilitators modified the instructional plan in delivering some topics. These modifications ranged from minor modifications (e.g., reducing duration of an activity) to major modifications (e.g., changing the original instructional plan completely). And (3) "Content"--facilitators ignored or added topics/activities during delivery (e.g., omission of the research findings during the presentation of the technical qualities of performance assessments.)

The most obvious impact of superficial modifications was the change in the schedule for implementing the program. For example, in tryout 1, there was not enough time to discuss some topics or carry out some activities on the last day. Information collected from direct observation revealed that

some of the process modifications led to discussions or raised questions that facilitators could not address properly. For example, on the first tryout at Site 1, facilitators were not able to close off discussions or to comment accurately on participants' accurate/inaccurate statements. At Site 2, process variations impacted program material. For example, by changing the instructional plan, facilitators had to change the "participants notebook". Finally, the evaluation found that the variations in the implementation of the program did not impact the achievement of the program's goals.

(b) Consistently across tryouts, the information delivered for certain topics was inaccurate (e.g., procedure-based scoring systems). Probably these topics were too complex for Facilitators to come to understand by reading the Facilitators' Manual. More emphasis should be paid to technical topics during facilitator training.

(c) Facilitators' knowledge and background were found to be a key to the success of program delivery. The Facilitators' knowledge about hands-on science instruction, their experience with the performance assessments used in the program (e.g., experience in administering and scoring performance assessments), their background in developing performance assessments, and their knowledge of the program content were all key factors in the quality of the delivery. For example, based on direct observation of delivery, we found that facilitators had difficulty identifying participants' misconceptions about performance assessments or answering "non-scripted questions" due to limited knowledge about and experience with hands-on instruction and performance assessments.

Direct observation also revealed that Facilitators tended to "know" only those parts of the manual they delivered; the rest of the content was not included in "their program." Facilitators at Site 1 displayed little

understanding of the agenda and the content to be presented for each succeeding session. For example, on occasion 1 at Site 1, one participant asked facilitators if they would have the opportunity to watch a video in which performance assessments were administered to a whole class. Facilitators said "no" even though there was a video on this issue that participants would see the next day. Facilitators had not read the whole manual or previewed the videos included in the program package.

As might be expected, this situation influenced the quality of the delivery. For example, facilitators, instead of closely monitoring small group discussions/activities, tended to use that time to read the next topic in the manual. Yet, hearing participants' discussions and viewing their performances would help facilitators discover misconceptions or problems in understanding.

(d) Observation of the delivery revealed that the certain characteristics of the participants made a difference in the type of questions, discussions, and even level of enthusiasm about the topics presented in the program. Participants from Site 2, already aware of the need for alternative assessments in their hands-on science curriculum, were more focused in their discussions and the questions they posed to facilitators, and clearly more enthusiastic.

Findings from the program delivery evaluation called for some changes on the TEP, specially after tryout 1: (a) The content of the program needed to highlight the importance of certain critical topics to the facilitators. (b) Background material from popular professional journals for teachers and administrators needed to be provided (e.g., special issues of Educational Leadership). Articles from research journals did not motivate facilitators to read about performance assessments. And most importantly, (c) facilitators

needed more thorough training on the content of the program before they could "own" it and deliver it successfully.

Program Materials. Major changes in the program materials were made after tryout 1. Three factors influenced these changes: facilitators' recommendations about the program, the variations observed during implementation, and an executive decision.

Facilitators made the following recommendations: the program needed more group discussion and more time for participants to "process" the information; the program included too much information, "less is more" according to one of the facilitators; the presentation of the technical qualities of performance assessments did not require as many charts (i.e., graphs) as those included; and the format of the manual needed improvement.

Information from the facilitator interviews conducted by one of the program developers revealed some reasons that facilitators had for modifications they made during the implementation. For example, they omitted some topics for two main reasons: (a) they considered the information irrelevant for teachers (e.g., research findings on the technical characteristics of performance assessments), or (b) time constraints. They included many new activities (e.g., "icebreakers", "sponge activities", "carousels", "prizes") because they wanted to give participants the opportunity to get up and move around and discuss ideas with each other, to keep them involved in the activities, and to provide positive feedback.

After the first tryout at Site 1, the principal investigator of the project made an "executive decision" to let program developers change the "format", not the "content" of the program. Because the content had already proven successful in achieving the program's goals, the developers' task was to enhance the "teacher-friendliness" of the program.

Discussions about changes in the instructional were the every day story in the project. Moreover, with changes in the plan came reviews and revisions of the content to insure that its accuracy was intact after modifications. Finally, a new version of the program material was tried out on the second occasion at Site 1, and at Site 2. Although the content remained almost the same, the instructional plan for delivery was substantially modified.

The new version of the program had the following characteristics: (a) Some topics were dropped from the program (e.g., two types of validity, curriculum sensitivity and discriminant validity), and others were reduced in scope (e.g., interrater reliability and intertask reliability). (See Appendix C for a list of the topics included in the program.) (b) New instructional activities were included to communicate content (e.g., mini-lectures; participant content checks.) (c) The instructional plan for delivering some topics was completely modified. For example, to motivate the review of the technical qualities of performance assessments prior to the selection exercise, the plan called for participants to improvise skits, in small groups, which depicted a specific technical quality (e.g., interrater reliability), act it out, and let the other participants guess the term depicted. (d) The content sequence was changed (see Appendix B.) (e) Characteristics of the Facilitators' Manual were changed (e.g., shaded boxes were used to indicate important points to make to the participants; mini-lectures were highlighted by a "box" and change in type font and size.) (f) The Participants' Notebook was modified based on changes in program content.

Reviews of the Facilitators' Manual during the modification period revealed new content inaccuracies as a result of the modifications (e.g., developers used inaccurate terms) and some instructional inadequacies (e.g.,

developers wanted participants to "fill-in the blanks" as a check on content knowledge, an activity that contradicted the hands-on philosophy of the program.)

Information from the Facilitators Questionnaire showed that, for facilitators, the TEP was effective in meeting its goals, and at an adequate level for the participants. In general, they thought that most of the segments were effective in meeting their particular objectives. However, all facilitators agreed that Segments 4, "Paper Towels Investigation", and Segment 5, "Introduction to Scoring Systems", were only "somewhat effective." Information collected from direct observation during the delivery corroborated this finding. Facilitators on all three tryouts had problems delivering these two segments.

Evaluation findings from the last two tryouts revealed that the "New Facilitator's Manual" still need improvement. Topics that proved to be consistently inaccurately delivered need to be revised to help facilitators deliver the content more accurately. Also, training need pay particular attention to these topics in the future.

Moreover, some of the "new activities" included in the modified manual should be presented as "optional" for facilitators. For example, according to facilitators from Site 2, some of the activities may not be appropriate for high school teachers (e.g., the improvisational skits.)

Finally, for facilitators to acquire the "whole" picture of the program, they needed to spend a considerable amount of time and effort studying and learning the material. There is no doubt that the amount of information contained in the Facilitators' Manual is a contributing factor. Facilitators felt overloaded with the information, which may have influenced the effort they put into "owning" the program. However, the commitment that facilitators

have to deliver a program with good quality is also an important factor. Facilitators at Site 2 were clearly more engaged with the delivery of the program. They met as a team at least four times before implementing the program to discuss content, modifications, and how the modifications impacted the delivery (e.g., schedule of the implementation, changes to the participants' notebook). These meetings lead facilitators to know better the characteristics of the program. In sum, delivery at Site 2 is a good example of the difference that facilitators' commitment can make.

Program Outcomes. The participants' self-reported knowledge about the topics addressed in the program and their opinions about the program were used as a source of information on outcomes.

The Self-Report Knowledge Inventory was administered to participants in a pretest-posttest design at each tryout. Designed to assess knowledge and skills acquired during the program, this instrument was somewhat unique. It had the appearance of a questionnaire and asked participants to indicate their understanding of the topics covered in the program.

To examine the differences in the knowledge and skills acquired by participants as a result of the program, a series of dependent t-tests for differences between pre- and posttest mean scores was carried out for each tryout. Table 6 presents the descriptive statistics and the reliability coefficients (i.e., internal consistency) at pre- and the posttest.

Insert Table 6 Here

Significant differences between the pre- and the posttest mean total scores were observed across tryouts ($t_{(29)} = 15.048$, $p < .01$; $t_{(16)} = 9.78$, $p < .01$; $t_{(19)} = 14.26$, $p < .01$, respectively.) Reliability coefficients for total scores were

high and roughly of the same magnitude at both pre- and posttest across the three tryouts. Patterns of differences between the pre- and posttest were similar to those observed on the last two in-house tryouts.

Based on participants report of their knowledge, the program was effective in achieving its goals despite program modifications made by the facilitators during the delivery.

Examples of the Evaluation Findings

In this section we present two concrete examples of evaluation findings. One addresses program delivery findings and the other content findings.

Program Delivery. To examine program delivery, data were collected on each of the topics in the program (see Appendix A & C). Findings on the delivery of one topic across the three tryouts are presented concisely in a flowchart (Figure 4). The shaded boxes symbolize the quality of delivery--the accuracy of content and adequacy of the plan for delivering it. The darker the box, the poorer the delivery quality. A light box indicates that the desired quality was achieved. The criteria used to shade the boxes based on the quality of the delivery is presented in Figure 5. Thick-line boxes refer to the evaluation findings; square boxes refer to the evaluator's recommendations, and arrows represent the time sequence.

Insert Figure 4 Here

Insert Figure 5 Here

As an example, consider the delivery of the topic, "Intertask Reliability," of performance assessments. Figure 4 presents the evaluation findings from direct observation bearing on the delivery of this topic across three tryouts.

This topic dealt with technical characteristics of performance assessments (i.e., reliability, validity, utility). Its purpose was to make participants aware of the importance of considering the consistency of students' performance-assessments scores across different tasks.

The evaluation on the first and the last tryouts showed that the quality of delivery was poor. Although the inaccuracies in the information delivered to participants were different, both reflected the fact that facilitators needed more information and better understanding of this topic in order to deliver it accurately. During the second tryout, the quality of the delivery improved. Two factors influenced this improvement. First, Facilitator 2 delivered this topic on both occasions. Second, on occasion 2, the facilitator discussed the topic with the Coordinator of Level I training before it was delivered. Both factors, the experience in delivering the topic and the discussion, may have helped.

Participants' perceptions were used to triangulate on the findings from direct observation. Participants in tryout 1 were asked, "What recommendations would you make about the organization and the content of this workshop that you think would help to improve it?" Some participants recommended dropping the topics related to statistical issues (e.g., "get rid of statistical information on validity and reliability"; "leave out the statistical lesson"). Furthermore, two participants wrote that it was clear that facilitators did not know enough about the statistical information.

Future training of other facilitators, then, should provide a more detailed explanation of the meaning of consistency across tasks as well as the relevance of this topic in the context of performance assessment.

Program Content. Evaluation findings on program material were based on careful review and revision of the Facilitators' Manual, Participants' Notebooks, and Transparencies. The criteria used to shade the boxes based on the quality of the content is presented in Figure 6.

Insert Figure 6 Here

Findings for the same topic, "Intertask Reliability of Performance Assessments," are presented in Figure 7. This topic is discussed in different parts of the content through different activities (see Appendix C). One of the relevant activities is the "Selection Exercise". Here participants have the opportunity to review and apply knowledge acquired during the program by selecting among four different performance assessments on electricity.

Insert Figure 7 Here

Based on the reviews, done by the evaluator and the Coordinator of the Level I training, the evaluation pointed out that the content presented to facilitators for discussing intertask reliability in the selection exercise was inaccurate. The impact of the evaluation findings was immediate and new and accurate information was included. It is important to mention that the iterative reviews and revisions to the program material did not correspond to

each iterative tryout. Many reviews and revisions were carried out before tryout 2 as a consequence of the modifications made to the program.

Future Dissemination Alternatives Based on the Evaluation of the Prototype TEP

Although the findings are encouraging, additional work is needed. For example, the length of the program has been a major concern. The difficulty of scheduling three-day workshops with teachers during the school year called for a variety of alternative schedules. Outside tryout 3 showed that the program could be delivered in two full days. The project has also developed two optional programs based on the three-day prototype program: A one-day workshop, and a three-hour workshop. Both have already been successfully implemented in different sites with project staff as facilitators (i.e., the principal investigators and the Coordinator of the Level I program).

A final dissemination concern focuses on the facilitators. It is important to remember that even though the outcomes indicated a positive program effect for these tryouts, direct observation revealed that facilitators delivered inaccurate information when discussing participants' questions or in talking in small/large group discussions. In the final analysis, facilitators' knowledge should be more than just the program content. They need a solid background in hands-on science and performance assessments. For example, facilitators need to know how performance are developed and how their psychometric characteristics are tested.

Unfortunately, this pilot test did not examine the "training program" for facilitators. Clearly improvements are needed. Nevertheless, information obtained from these tryouts should help to develop a set of training guidelines. For example, it is clear that first, facilitators should

experience the program as participants. Moreover, training should focus on those topics that the evaluation has revealed to be particularly difficult for facilitators. Training should also provide facilitators with the opportunity to administer, score, and interpret performance assessments so that they feel comfortable in explaining ideas and answering questions. Finally, training should also incorporate information about performance assessments from journal articles and magazines.

Because only two sites were used to tryout the program it is not possible to delimit the range of settings most appropriate for implementing the program. However, information accumulated through the formative evaluation process suggests that settings in which the school district already recognizes that alternative assessment is a crucial part of hands-on science instruction are the most suitable sites in which to implement the program. This does not mean that the program cannot be implemented in settings only planning or starting to change toward a hands-on science instructional approach. It only means that the benefits from the TEP will be more obvious for those participants who are already aware of the necessity of this type of assessment in order to be congruent with new forms of instruction.

The program is also suitable for those settings in which the district is beginning a system of development and implementation of performance assessments. In these settings (e.g., resource centers), this TEP can be seen as the first step before moving to a TEP that can help them develop performance assessments.

Based on the study findings one alternative that has been considered as a way to solve the problem of "facilitator expertise" is to transfer the program to private-sector and government-funded organizations that will take responsibility for training teachers and administrators. This means that the

prototype TEP will be made available to "skilled users" at research and development centers, and to a few school districts whose staff have received extensive training in its use.

Conclusions

A central task during the evaluation of a prototype program is to study how delivery and outcomes vary from site to site (e.g., Cronbach, 1982). Evaluation of prototype programs should provide information about what to expect and what to do when the TEP program becomes fully operational.

Evaluation findings across three tryouts showed that the program was successfully implemented in outside sites. In other words, it is "robust" (Berk & Rossi, 1990). It produced similar results with different facilitators despite the variations in the ways the program was implemented. However, if success is defined in terms of the accuracy of all the information delivered during the implementation of the program, the program was less successful.

Still, evaluation results encourage the dissemination of the program at different sites. Very well trained facilitators with extensive knowledge of hands-on instruction and performance assessments will help to deliver an accurate and effective TEP.

The approach used for the evaluation study proved to be helpful in achieving the evaluation goals. The iterative process provided the opportunity to test modifications made to the program on the basis of the evaluation findings and to accumulate knowledge about the program. This knowledge contributed to a better understanding of the TEP by pinpointing the conditions needed for achieving the program's goals.

References

Berk, R. A. & Rossi, P. H. (1990). Thinking about program evaluation. Newbury Park: Sage Publications.

Chinien, C. & Hlynka, D. (1993). Formative evaluation of prototype products: From expert to connoisseur. Educational and Training Technology International, 30(1), 60-66.

Cronbach, L. J. (1982). Designing evaluations of educational and social programs. San Francisco: Jossey-Bass Publishers.

Cronbach, L. J. & Associates (1980). Toward reform of program evaluation. Aims, methods, and institutional arrangements. San Francisco: Jossey-Bass Publishers.

Ellis, J. D. & Goulding, P. G. (1991). Evaluation of ENLIST Micros III: Teacher Centers for improving the use of microcomputers in science instruction. An Interim Report to the National Science Foundation.

Ellis, J. D. & Kuerbis, P. J. (1991). A curriculum for preparing science teachers to use microcomputers. School Science and Mathematics, 91(6), 247-254.

Fitzsimmons, S. J., Carlson, K., Burnham, R., Heinig, S., & Stoner, D. (in preparation). A study of NSF Teacher Enhancement Program principal investigators: 1984-1989 (Draft). Cambridge, MA: Abt Associates Inc.

Gayford, C. (1987). Biotechnology 13-18: In-service training for teachers. Journal of Biological Education, 21(4), 281-287.

Guskey, T. & Sparks, D. (1991). What to consider when evaluating staff development. Educational Leadership, 49(3), 73-76.

Hurd, P. H. (1986). Perspectives for the reform of science education. Phi Delta Kappan, 67(5), 353-358.

Joyce, B. (1988). Training research and preservice teacher education: a reconsideration. Journal of Teacher Education, 39(5), 32-36.

Joyce, B. & Showers, B. (1980). Improving inservice training; The messages of research. Educational Leadership, 37(5), 379-385.

Knapp, M. S., Shields, P. M., ST. John, M., Zucker, A. A., & Sterns, M. S. (1988). An approach to assessing initiatives in science education: Summary Report: Recommendations to the National Science Foundation. SRI Project No. 089. Stanford, CA: SRI International.

Kulm, G. and Stuessy, C. (1991). Assessment in science and mathematics education reform. In G. Kulm and S. Malcom (Eds.). Science assessment in the service of reform (pp. 71-87). American Association for the Advancement of Science.

Nunnally, J. C. (1970). Introduction to psychological measurement. New York: McGraw-Hill.

Patton, M. Q. (1984). Data collection: Options, strategies, and cautions. In L. Rutman (Ed.). Evaluation research methods: A basic guide (2nd. ed.) (pp. 39-63). Beverly Hills: Sage Publications.

Raizen, S. A., Baron, J. B., Champagne, A. B., Haertel, E., Mullis, I. V. S., & Oakes, J. (1989). Assessment in science education: The middle years. Washington, DC: The National Center for Improving Science Education.

Ruiz-Primo, M. A. (1994). Formative Evaluation of Teacher Enhancement Programs: An Approach and Case Study. Unpublished doctoral dissertation, University of California, Santa Barbara.

Ruiz-Primo, M. A., Shavelson, R. J. & Baxter, G. P. (1993). An Approach to Formative Evaluation for Teacher Enhancement Programs. Paper presented at the American Educational Research Association Annual Meeting. Atlanta, Georgia, 1993.

Scriven, M. (1967). Goals of evaluation versus roles of evaluation: Formative and summative evaluation. In R. W. Tyler, R. M. Gagné, & M. Scriven (Eds.) Perspectives of curriculum evaluation. AERA Monographs on Curriculum Evaluation (No. 1). Chicago: Rand McNally.

Scriven, M. (1991a). Evaluation Thesaurus (4th ed.), Newbury Park, CA: Sage Publications.

Scriven, M. (1991b). Beyond formative and summative evaluation. In M. McLaughlin & D. C. Phillips (Eds.). Evaluation and Education: At quarter century (pp 19-64). Ninetieth Yearbook of National Society for the Study of Education. Chicago, IL: The University of Chicago Press.

Shadish, W. R. Jr., Cook, T. D., & Leviton, L. C. (1991). Foundations of program evaluation. Theories of practice. Newbury Park: Sage Publications.

Shavelson, R. J. & Baxter, G. P. (1990). Transfer new assessment technologies. to teachers and other educators. National Science Foundation Grant No. TEP 90-55443.

Shavelson, R. J., Carey, N. B., and Webb, N. M. (1990). Indicators of science achievement: Options for a powerful policy instrument. Phi Delta Kappan, 71(9), 692-697.

Shavelson, R. J., Copeland, W., Baxter, G. P., Decker, D. L, & Ruiz-Primo, M. A. (1994). Inservice education models for enhancing the teaching of science. In S. Fitzsimmons and L. C. Kerpelman (Eds.) Teacher enhancement for elementary and secondary science and mathematics: Status, issues, and problems. Washington, D.C.: National Science Foundation. Division of Research, Evaluation and Dissemination. Directorate for Education and Human Resources.

Sparks, D. & Loucks-Horsley, S. (1990). Models of staff development. In W. R. Houston, M. Haberman, & J. Sikula (Eds.) Handbook of research on teacher education (pp. 235-250). New York: Macmilan Publishing Company.

Tamir, P. & Amir, R. (1981). Retrospective curriculum evaluation: An approach to the evaluation of long term effects. Curriculum Inquiry, 11(3), 259-278.

Weiss, C. H. (1972). Evaluation research. Englewood Cliffs, NJ: Prentice-Hall.

Young, D. B. & Tamir, P. (1977). Finding out what students know. The Science Teacher, 44, 27-28.

TABLE 1

Characterization of an Approach to Formative Evaluation of TEPs.

TEP PROGRAM		EVALUATION PROCESS			EVALUATOR	
Stage of Maturity of the Program	Program's Components	Iterative Process: Conditions	Stage of Formative Evaluation:	Diversity of Methods of Evaluation	Adaptable Role	Subject-Matter Expert
Planned Program	Context Goals Materials Conditions of Delivery	Reviews-Revisions of the program before it is tried out.	In-house Reviews: Information about the accuracy and feasibility of the program.	Informal evidence (e.g., comments on content and possible difficulties).		
Experimental Program	Context Goals Materials Delivery Outcomes	Tryouts of the program with in-house staff and typical consumers. Revisions and reviews are also appropriate.	In-house Tryouts: Information about program operation: Variations and characteristics associated with more or less successful components.	Quantitative (e.g., small studies using quasi experimental or randomized designs) and qualitative methods (e.g., case studies)		
Prototype Program	Context Goals Materials Delivery Outcomes	Tryouts of the program in different sites with similar conditions to those proposed for operation. Revisions. Reviews are less necessary but may be appropriate too.	Outside Tryouts: Information on differences in delivery and effects from site to site, possible problems, and costs of implementation.	Quantitative and qualitative methods using research designs for estimating effectiveness are highly recommended		

TABLE 2

Design of the Formative Evaluation of the Program to Transfer Performance Assessment Technology

			Tryouts				
			Bef-	Delivery			Af-
Focus of the Evaluation	Source of Information	Method of Data Collection	ore	Pre	Dur ing	Post	ter
Program Delivery	Evaluator	Direct Observation			X		
	Participants	Opinion Questionnaire				X	
Program Material	Documents	Review of documents	X				
	Participants	Opinion Questionnaire				X	
	Facilitators	Questionnaire					X
Program Outcomes	Participants	Self-report knowledge and skills questionnaire		X		X	
		Opinion Questionnaire				X	

TABLE 3

General Characteristics of the Three Outside Tryouts of the Level I Training

ASPECTS	OUTSIDE TRYOUTS		
	SITE 1		SITE 2
	1	2	3
Facilitators' Characteristics	Both were Elementary Teachers and Staff Developers from the County School Office	Same Facilitators as Occasion 1	Two Resource Teachers from the District and One Elementary Teacher
Participants' Characteristics	Elementary Teachers	Elementary Teachers	Elementary, High School, and Resource Teachers
Participants' Incentives	Ps paid for the workshop and got one unit credit	Ps paid for the workshop and got one unit credit	None
Organization of the Program in Days	3 in a row	2 evenings and one full-day	2 full-days
Methods of Data Collection			
• Facilitator's Critique Questionnaire		√	√
• Direct Observation	√	√	√
• Review of Documents	√	√	√
• Self-Report Knowledge Inventory	√	√	√
• Opinion Questionnaire	√	√	√
• Descriptive Information Questionnaire	√	√	√

TABLE 4

Characteristics of the Facilitators By Site

CHARACTERISTICS	OUTSIDE TRYOUTS				
	SITE 1		SITE 2		
	F[a]1	F2	F1	F2	F3
Years of Teaching Experience					
Elementary	7		28		6
Junior High		6		3	
High School		10		20	
Years of Science Hands-On Teaching					
Elementary	7		26		6
Junior High		6			
High School		10		20	
Years as Trainer of Science Teachers	4	2	9	6	3
Undergraduate Major					
Education	√				√
Science		√		√	
Other			√		
Advanced Degree					
MA Education	√		√	√	√
MA Science		√			
MA Other					
Ph. D. Education					√
Experience with Hands-On Science Teaching					
Novice					
User		√		√	NA[b]
Expert	√		√		
Experience with Science Performance Assessments					
Novice			√		
User	√	√		√	√
Expert					

[a] F = Facilitator.
[b] NA = No Answer

TABLE 5

Characteristics of the Participants Across the Outside Tryouts

CHARACTERISTICS	OUTSIDE TRYOUTS		
	SITE 1		SITE 2
	1	2	3
Number of Participants	38	20	22
Mean Years of			
Elementary Teaching	13.29	10.93	6.09
Junior High Teaching	0.93	.50	7.52
High School Teaching	0.38	.96	9.56
Mean Years of Hands-on			
Elementary Teaching	6.9	7.81	5.57
Junior High Teaching	0.45	.06	5.38
High School Teaching	0.38	0.0	0.19
Undergraduate Major			
Education	9	2	9
Science	1	1	11
Other	21	13	2
Advanced Degree			
MA Education	7	4	10
MA Science	0	0	5
MA Other	4	1	1
Ph. D. Education	0	0	1

TABLE 6
Statistics for the Pre- and the Posttest Total Score

	1		2		3	
	Pre	Post	Pre	Post	Pre	Post
nparticipants	38		20		22	
Maximum	48		64		68	
Mean	19.20	38.22*	30.60	51.82*	31.57	56.47*
S. D.	6.40	5.15	6.26	7.08	8.72	7.52
Reliability	.91	.86	.87	.94	.95	.93

* Significant difference between Pre- and Posttest (a = .001).

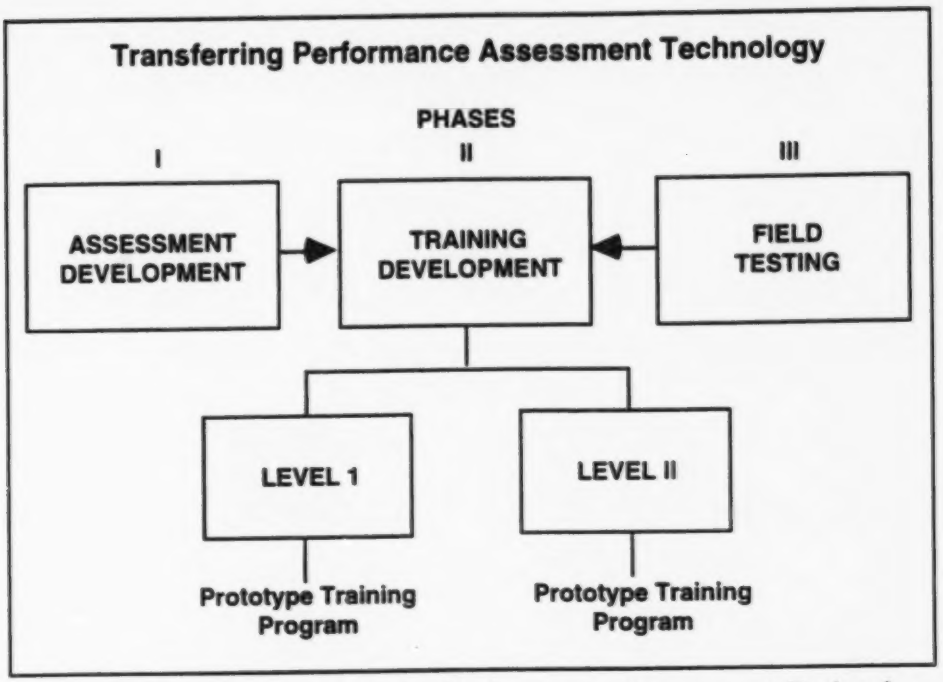

Figure 1. A Project to Transfer Performance Assessment Technology to Teachers and Other Educators.

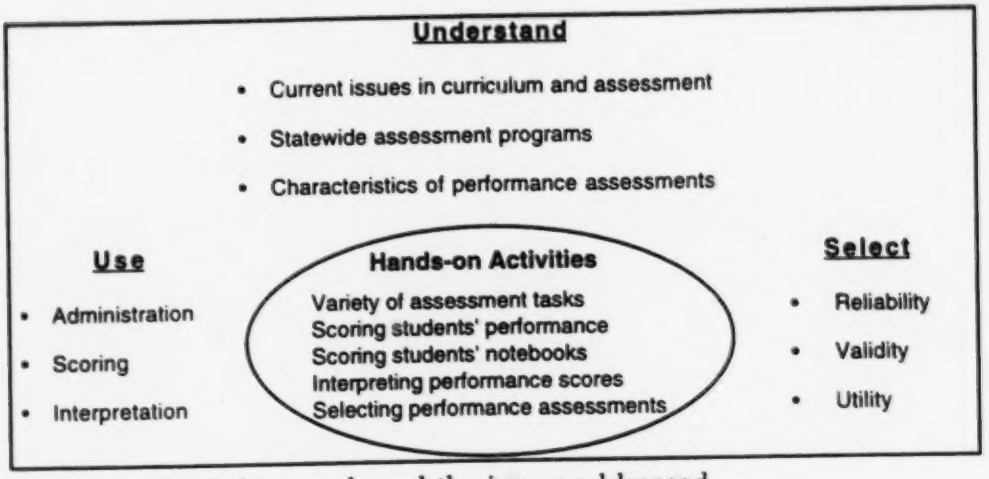

Figure 2. Workshop goals and the issues addressed.

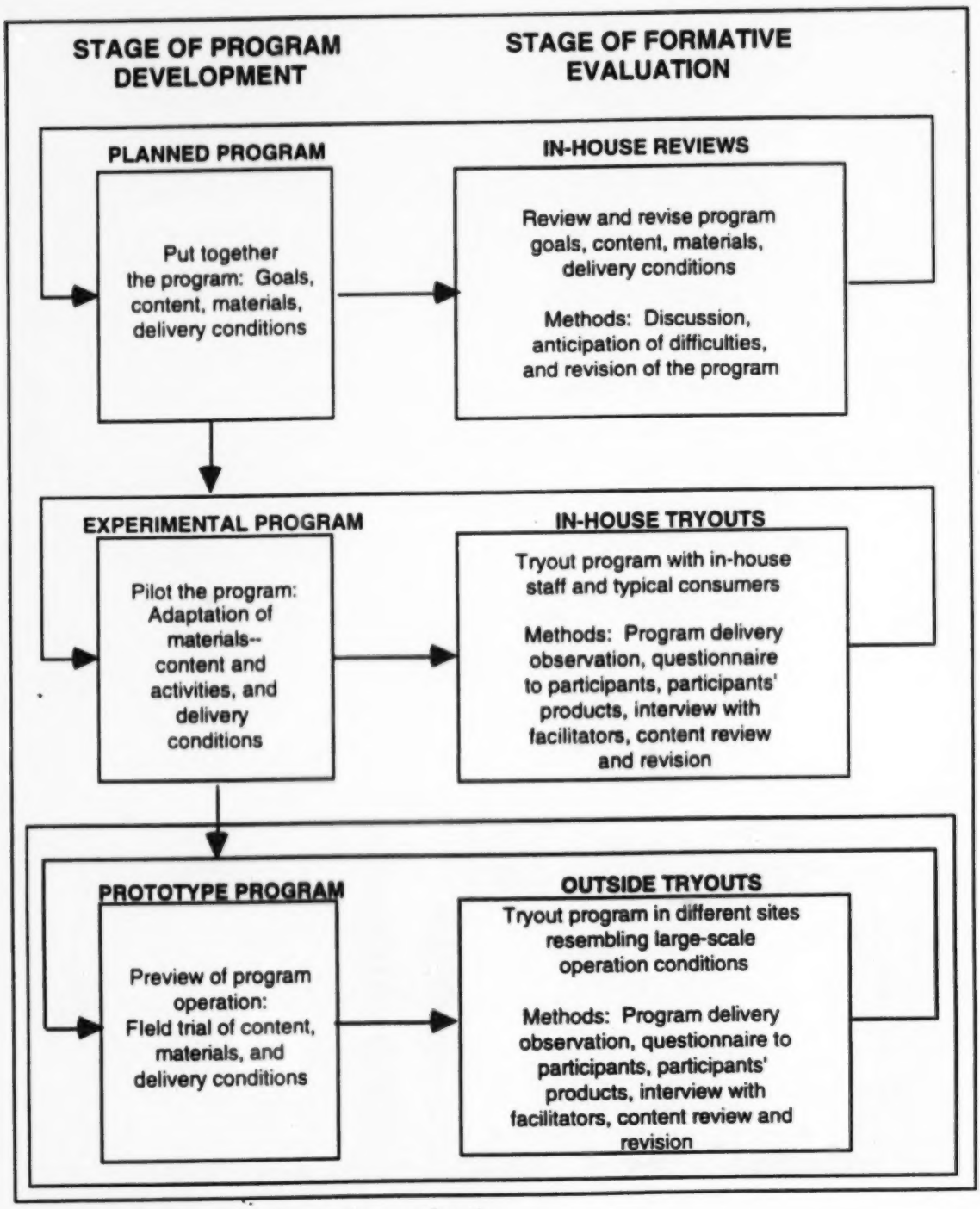
STAGE OF PROGRAM DEVELOPMENT
STAGE OF FORMATIVE EVALUATION
PLANNED PROGRAM
Put together the program: Goals, content, materials, delivery conditions
IN-HOUSE REVIEWS
Review and revise program goals, content, materials, delivery conditions
Methods: Discussion, anticipation of difficulties, and revision of the program
EXPERIMENTAL PROGRAM
Pilot the program: Adaptation of materials-- content and activities, and delivery conditions
IN-HOUSE TRYOUTS
Tryout program with in-house staff and typical consumers
Methods: Program delivery observation, questionnaire to participants, participants' products, interview with facilitators, content review and revision
PROTOTYPE PROGRAM
Preview of program operation: Field trial of content, materials, and delivery conditions
OUTSIDE TRYOUTS
Tryout program in different sites resembling large-scale operation conditions
Methods: Program delivery observation, questionnaire to participants, participants' products, interview with facilitators, content review and revision

Figure 3. Strategy for formative evaluation.

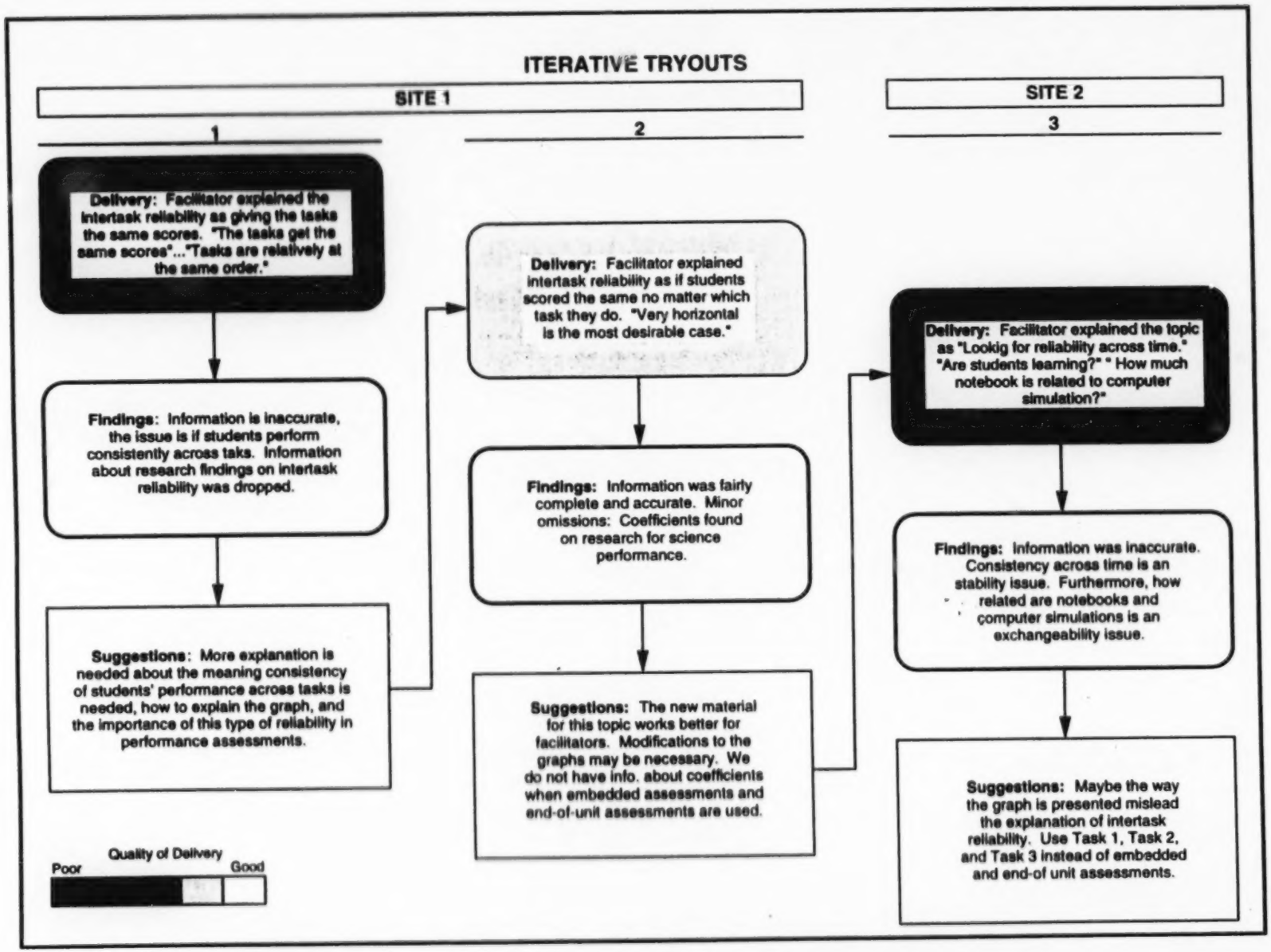
ITERATIVE TRYOUTS
SITE 1
SITE 2
1
2
3
Delivery: Facilitator explained the intertask reliability as giving the tasks the same scores. "The tasks get the same scores"..."Tasks are relatively at the same order."
Findings: Information is inaccurate, the issue is if students perform consistently across taks. Information about research findings on intertask reliability was dropped.
Suggestions: More explanation is needed about the meaning consistency of students' performance across tasks is needed, how to explain the graph, and the importance of this type of reliability in performance assessments.
Delivery: Facilitator explained intertask reliability as if students scored the same no matter which task they do. "Very horizontal is the most desirable case."
Findings: Information was fairly complete and accurate. Minor omissions: Coefficients found on research for science performance.
Suggestions: The new material for this topic works better for facilitators. Modifications to the graphs may be necessary. We do not have info. about coefficients when embedded assessments and end-of-unit assessments are used.
Delivery: Facilitator explained the topic as "Lookig for reliability across time." "Are students learning?" " How much notebook is related to computer simulation?"
Findings: Information was inaccurate. Consistency across time is an stability issue. Furthermore, how related are notebooks and computer simulations is an exchangeability issue.
Suggestions: Maybe the way the graph is presented mislead the explanation of intertask reliability. Use Task 1, Task 2, and Task 3 instead of embedded and end-of unit assessments.
Quality of Delivery
Poor
Good

Figure 4. Delivery of the topic "Intertask Reliability of Performance Assessments."

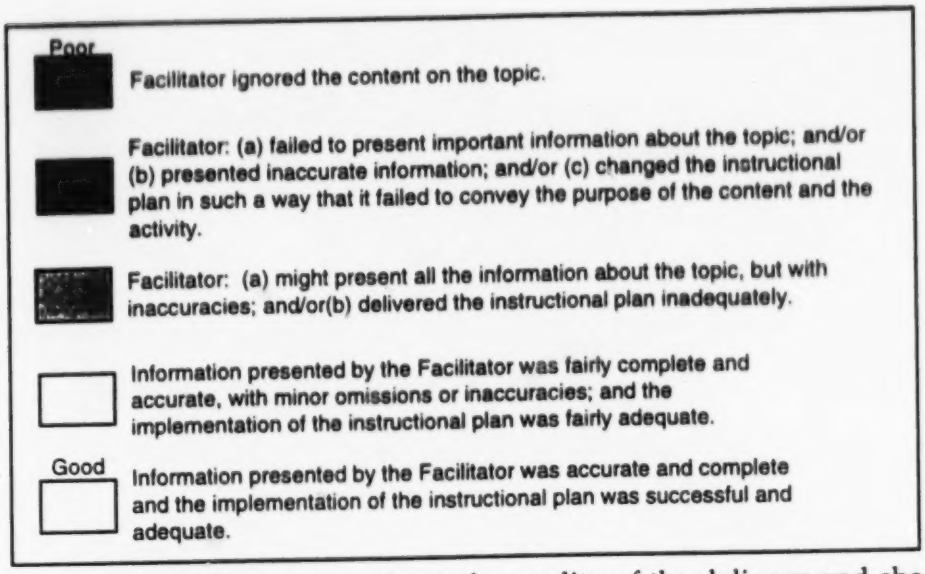
Poor
Facilitator ignored the content on the topic.
Facilitator: (a) failed to present important information about the topic; and/or (b) presented inaccurate information; and/or (c) changed the instructional plan in such a way that it failed to convey the purpose of the content and the activity.
Facilitator: (a) might present all the information about the topic, but with inaccuracies; and/or(b) delivered the instructional plan inadequately.
Information presented by the Facilitator was fairly complete and accurate, with minor omissions or inaccuracies; and the implementation of the instructional plan was fairly adequate.
Good
Information presented by the Facilitator was accurate and complete and the implementation of the instructional plan was successful and adequate.

Figure 5. Criteria used to evaluate the quality of the delivery and shade the boxes in the flowcharts.

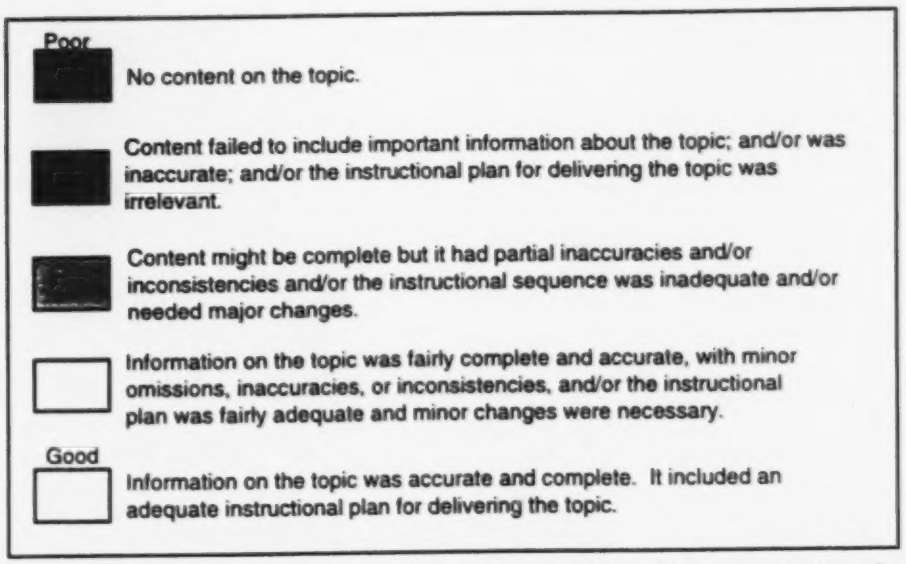
Poor
No content on the topic.
Content failed to include important information about the topic; and/or was inaccurate; and/or the instructional plan for delivering the topic was irrelevant.
Content might be complete but it had partial inaccuracies and/or inconsistencies and/or the instructional sequence was inadequate and/or needed major changes.
Information on the topic was fairly complete and accurate, with minor omissions, inaccuracies, or inconsistencies, and/or the instructional plan was fairly adequate and minor changes were necessary.
Good
Information on the topic was accurate and complete. It included an adequate instructional plan for delivering the topic.

Figure 6. Criteria used to evaluate the quality of the content and shade the boxes in the flowcharts.

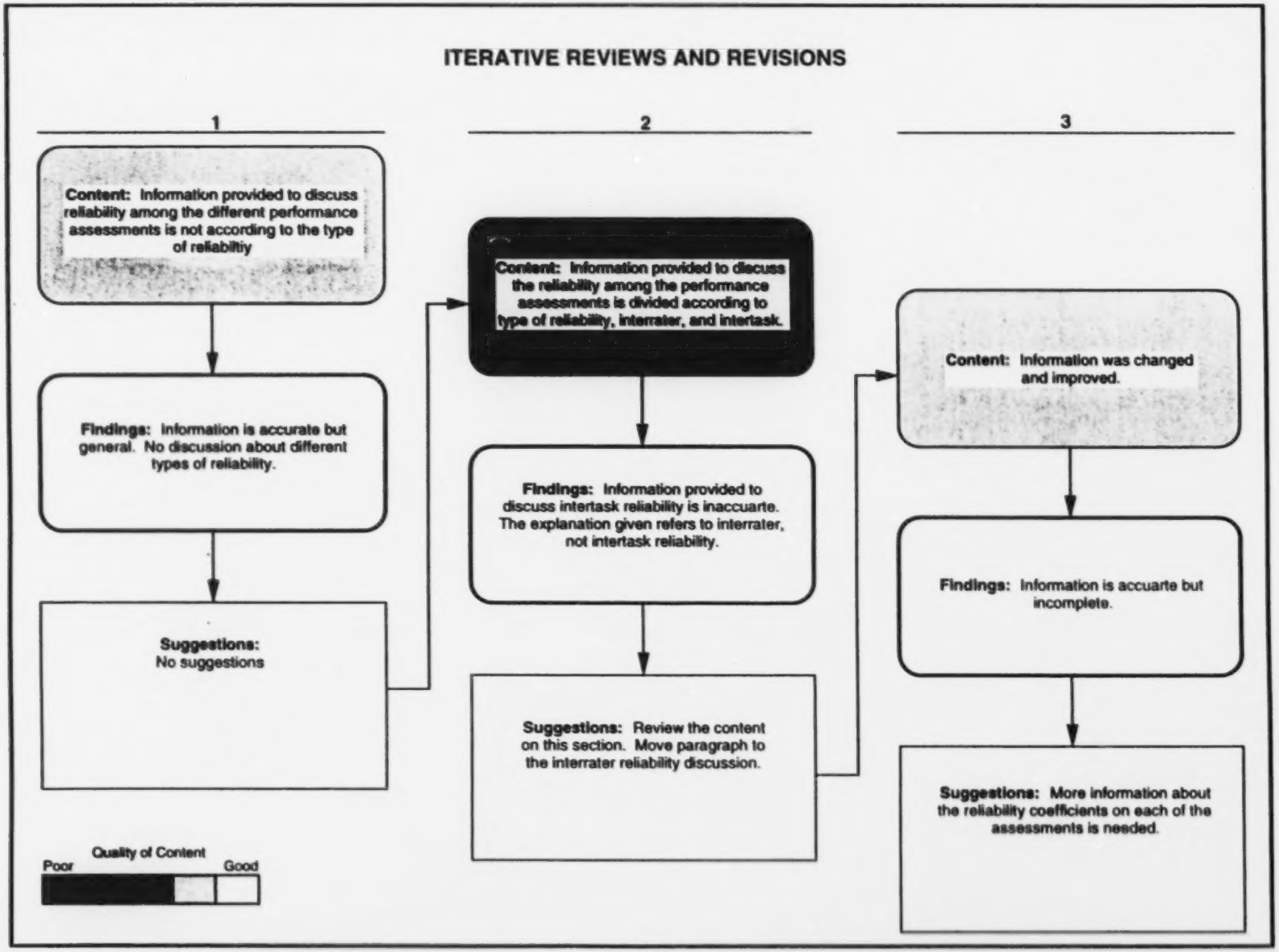
ITERATIVE REVIEWS AND REVISIONS
1
2
3
Content: Information provided to discuss reliability among the different performance assessments is not according to the type of reliabiltiy
Findings: Information is accurate but general. No discussion about different types of reliability.
Suggestions: No suggestions
Content: Information provided to discuss the reliability among the performance assessments is divided according to type of reliability, interrater, and intertask.
Findings: Information provided to discuss intertask reliability is inaccuarte. The explanation given refers to interrater, not intertask reliability.
Suggestions: Review the content on this section. Move paragraph to the interrater reliability discussion.
Content: Information was changed and improved.
Findings: Information is accuarte but incomplete.
Suggestions: More information about the reliability coefficients on each of the assessments is needed.
Quality of Content
Poor
Good

Figure 7. Quality of the content of the topic "Intertask Reliability" provided in the activity "Selection of Performance Assessments."

APPENDIX A
Topics and Issues Addressed in the Program by Goal

UNDERSTANDING

Current Issues in Curriculum and Assessments

1. Current assessments practices
   - Purposes of assessments
2. Current curriculum and assessment reform
   - Characteristics of hands-on instruction
   - Hands-on instruction and performance assessments as the two sides of the same coin

Statewide Assessment Program

3. California Assessment Program--An example of a large-scale assessment program
   - CAP assumptions
   - CAP information sources

Characteristics of Performance Assessments

4. Characterization of PA
   - Definition of assessment and performance assessments
   - Differences between traditional multiple-choice tests and performance assessments
   - Variability of performance
5. Different types of tasks
   - Variety of experiences with performance assessments
   - Comparative investigations. Example used: Paper Towels Investigation
   - Component identification Example used: Mystery Powders Investigation
   - Classification tasks Example used: Leaves Task
6. Different types of scoring forms
   - Analytic scoring forms
     - Procedure-based Example used: Paper Towels scoring form
     - Evidence-based Example used: Mystery Powders scoring form
   - Holistic scoring forms
     - Rubric Example used: Leaves rubric

USE

Administration of Performance Assessments

7. Management techniques to administer performance assessments in the classroom
   - Organization of materials
   - Help from students
   - Observation during assessments
8. Embedded and End-of-Unit assessments Example: Mystery Powders embedded assessment # 2.

Scoring Performance Assessments

9. Range of alternative assessments
   - Array of performance assessments
   - Direct observation of students' performance
   - Notebooks as surrogates of direct observation
     - exchangeability
10. Practicing scoring students' performance
    - Scoring from direct observation. Example: Scoring Paper Towels (videos)
    - Scoring students' notebooks Example: Paper Towels, Mystery Powders, and Leaves

Interpreting Performance Assessments

11. Interpreting performance assessments scores
    - Summary of scores
    - Patterns of scores
    - Interpretation of patterns of scores

SELECTION

Reliability of Performance Assessments

12. Interrater reliability of performance assessments
    - High and low reliability coefficients
    - Research findings about interrater reliability
13. Intertask reliability of performance assessments
    - Research findings about intertask reliability

Validity of Performance Assessments

14. Criterion validity
    - Sensibility of performance assessments to different types of curriculum
      - Research findings about curriculum sensitivity
    - Comparison between multiple-choice traditional tests and performance assessments
      - Research findings about discriminant validity
    - Comparison between performance assessments and aptitude tests and multiple-choice and aptitude tests
      - Research findings about these correlations
15. Content validity
    - Overlapping of performance assessments and teaching unit
    - Logical judgment as evidence for content validity

Utility of Performance Assessments

16. Usefulness of scores
    - Utility of scores to evaluate students' performance
    - Utility of scores for curriculum monitoring
17. Easiness of the assessment in the classroom and the scoring form
18. Cost-effectiveness of the assessment
    - Time and cost

APPENDIX B

Sequence and Organization of the Program Content of the Prototype Program Across Tryouts 1, 2 and 3

TRYOUT 1	TRYOUT 2-3
1. Introductions and Overview • Initial introductions • Expectations and Pretest • Workshop goals: Characteristics, use, and selection of performance assessments • Review of agenda	**1. Introductions and Overview** • Workshop opening • Workshop goals • Participants' expectations • Facilitators' expectations • Current curriculum reform in science education * pre- and post-reform • Current assessment reform in science education * pre-and post-reform • Workshop format
2. Hands-on Experience with Traditional and Performance Assessments • Working with assessments * Multiple-Choice Tests Discussion of the tasks * Performance Assessment Circus • Current curriculum reform * characteristics of hands-on science instruction • Needed reform in assessment * two sides of same coin	**2. Experience with Assessments** • Introduction to assessments • Multiple-choice tests • Variety of performance assessments • Characteristics of performance assessments * performance task * performance product * scoring system • Management techniques in the classroom
3. Assessment and Performance Assessment • Current assessment practices * four purposes of assessment * definition of assessment * definition of performance assessment * range of performance assessments - direct observation - notebooks - computer simulation - short-answer questions - new multiple-choice items	**3. Assessments in Science Education** • Introduction to varieties of assessments * Mini-Lecture: Variety of assessments - portfolios - performance assessments methods - direct observation - assessment notebooks - computer simulation - new paper-and-pencil tests - exchangeability • Content check
4. Hands-on Experience with a Performance Assessment • Observation and scoring without scoring system • Discussion of results * variability in performance * variability in scoring	**4. Paper Towels Investigation** • Observation and scoring of Paper Towels Investigation • Score summary form • Small group discussion • Large group discussion • Scoring performance * Mini-Lecture: Scoring performance - variability in performance - variability of observers' scores - interrater reliability • Content check • Video of students conducting paper towels investigation

APPENDIX B (Continued)

TRYOUT 1	TRYOUT 2-3
5. Introduction to Scoring Systems • Introduction to a procedure-based scoring system for Paper Towels • Presentation of procedure-based scoring of Paper Towels * description and judgment • Practice scoring from a student's notebook	**5. Introduction to Scoring Systems** • Scoring direct observation of the Paper Towels task * Mini-Lecture: Scoring form for Paper Towels - scoring from direct observation - method for getting the towel wet - saturation - determining result - care in saturation and/or measuring - correct result - assigning a grade • Scoring the Paper Towels assessment notebook * Mini-Lecture: Assessment notebooks and scoring performance - assessment notebook - Miguel's notebook - method for getting the towel wet - saturation - determine result - care in saturation and/or measurement - correct result - assigning grade • Discussion on the scoring of assessment notebooks • Scoring Cecilia's assessment notebooks • Exchangeability of notebooks as surrogates of direct observation * Mini-Lecture: Exchangeability - notebooks as surrogates of direct observation - exchangeability - exchangeability across assessment methods
6. Technical Quality of Performance Assessments • Judging performance assessments • Criteria * reliability - interrater reliability research findings - intertask reliability research findings * validity - content validity research findings - curriculum sensitivity research findings - discriminant validity research findings * utility • Example in technical characteristics of performance assessments	**6. Mystery Powders Investigation** • Description of the Mystery Powders Unit * Mini-Lecture: Mystery Powders Unit - general description of Mystery Powders Unit - assessments in the Mystery Powders Unit • Performing the Mystery Powders embedded assessment task • Discussion of the Mystery Powders embedded assessment task * Mini-Lecture: Content validity - content validity • A scoring system for Mystery Powders Assessment Notebooks • Scoring students' mystery powders assessment notebook * Mini-Lecture: Steps in scoring - what's inside the bag - observing tests - quality of evidence score - determining total scores. - additional issues

APPENDIX B (Continued)

TRYOUT 1	TRYOUT 2-3

	• Additional practice scoring with Sam's assessment notebook
	• Introduction to interpreting performance assessments
	• Step 1: Compile a summary of individual scores
	• Step 2: Examine the summary form
	• The Placement of assessment in teaching units
	  * Mini-Lecture: Assessments in the Mystery Powders Unit
	  - general review of mystery powders assessment
	  - embedded assessment
	  - end-of-unit assessment

7. Hands-On Experience with Embedded Performance Assessment
• Description of the Mystery Powders Unit
• Performing the mystery powders embedded assessment task
• Discussing the task
• Scoring students' mystery powders notebooks
  * explanation of mystery powders scoring form
• Interpreting performance assessments
• Embedded and end-of-unit assessments
• Genres of performance assessments
  * comparison
  * decomposition
  * taxonomy
  * description
	7. Important Qualities of Performance Assessments
	• Qualities of performance assessments
	• Exploring qualities of performance assessments
	• Application to the Mystery Powders assessment
	• Reliability as a quality of performance assessments
	  * Mini-Lecture: Reliability
	  - definition
	  - interrater reliability
	  - intertask reliability
	  - research findings and their implications
	• Content Validity as a quality of performance assessments
	  * Mini-Lecture: Content validity
	  - definition of content validity
	  - questions teachers might ask about content validity
	  - research findings and their implications
	• Utility as a quality of performance assessment
	  * Mini-Lecture: Utility and practicality
	  - definition of utility
	  - questions teachers might ask about utility
	  - research findings on utility
	  - definition of practicality
	  - questions to be asked in determining practicality
	  - research findings on practicality

APPENDIX B (Continued)

TRYOUT 1	TRYOUT 2-3

8. Hands-On Experience with Another Performance Assessment
• Performing the CAP assessment task: Leaves
• Discussion of this new genre of assessment
• Introduction of a Holistic Scoring System: Discussion of the concept "Rubric"
  * development process of the rubric
• Scoring notebooks using a rubric
• Some insights into the future of large-scale performance assessment--The California case
• A closer look at the 1992 CAP performance assessment in science
	8. Bugs Investigation
	• Judging performance assessments
	• The bugs investigation
	  * Mini-Lecture: Bugs investigation
	  - description of mealworms unit
	  - description of bugs performance task
	  - description of bugs performance product
	  - description of bugs scoring system
	  - research of bugs investigation
	  - interrater reliability
	  - intertask reliability
	  - content validity
	  - exchangeability
	  - utility
	  - practicality

9. Selection of Performance Assessment
• Criteria for judging PA
  * reliability review and questions
  * validity review and questions
  * utility review and questions
• Selecting performance assessments
• Discussion of selection exercise
• Recommendations
• Conclusion to the Workshop
	9. Rubric Scoring
	• Current science assessment practice
	• The Leaves assessment
	• Introduction of a Rubric scoring system
	  - definition of rubric
	  - development of a rubric
	• Scoring Jacob's notebook using a rubric
	• Scoring Leticia's notebook using a rubric
	• General discussion on using a rubric
	• Distinctions among scoring systems
	  * Mini-Lecture: Distinctions among scoring systems
	  - additional distinctions among scoring systems
	  - scoring systems
	
	10. Selection of Performance Assessments
	• Management of performance assessments
	• Criteria for judging performance assessments or "Afternoon at the Improv"
	• Selecting performance assessments
	• Discussion of selection activity
	
	11. Review and Closing
	• Creation of visual representation of terms and concepts
	• Closing of workshop

APPENDIX C

Topics and Issues Addressed in the "New" Prototype Program by Goal

UNDERSTANDING	USE	SELECTION
Current Issues in Curriculum and Assessments 1. Current curriculum and assessment reform • Characteristics of science instruction pre- and post-science reform • Science Assessment pre- and post-science reform **Characteristics of Performance Assessments** 2. Performance Assessments Components • Performance Task • Performance Product • Scoring Form 3. Different types of tasks • Variety of experiences with performance assessments • Comparative investigations. Example used: Paper Towels Investigation • Component identification Example used: Mystery Powders Investigation • Classification tasks Example used: Leaves Task 4. Different types of scoring forms • Analytic scoring forms - Procedure-based Example used: Paper Towels scoring form - Evidence-based Example used: Mystery Powders scoring form • Holistic scoring forms - Rubric Example used: Leaves rubric 5. Performance Assessments Methods • Direct Observation • Notebooks • Computer Simulation • New Paper-and-Pencil Tests 6. Embedded and End-of-Unit Performance Assessments • Example: Mystery Powders Embedded Assessment #2	**Administration of Performance Assessments** 7. Management techniques to administer performance assessments in the classroom **Scoring Performance Assessments** 8. Practicing scoring students' performance • Scoring from direct observation. Example: Scoring Paper Towels (videos) • Scoring students' notebooks Example: Paper Towels, Mystery Powders, and Leaves **Interpreting Performance Assessments** 9. Strategy to Interpret Performance Assessments Scores • Summary of scores • Patterns of scores • Interpretation of patterns of scores Example: Interpreting Mystery Powders #2	**Reliability of Performance Assessments** 10. Interrater reliability of performance assessments • High and low reliability coefficients • Research findings about interrater reliability 11. Intertask reliability of performance assessments • Research findings about intertask reliability **Validity of Performance Assessments** 12. Content validity • Overlapping of performance assessments and teaching unit • Logical judgment as evidence for content validity **Exchangeability of Performance Assessments** 13. Exchangeability of: • Notebooks to direct observation • Computer simulation to direct observation • New paper-and-pencil to direct observation **Utility of Performance Assessments** 14. Usefulness of scores • Utility of scores to evaluate students' performance • Utility of scores for curriculum monitoring **Practicality of Performance Assessments** 15. Easiness of the assessment in the classroom and the scoring form 16. Cost-effectiveness of the assessment • Time and cost